# DATOS ENLAZADOS PARA VOCABULARIOS ABIERTOS Y MARCO GENERAL DE *HIVE*

**Eva Méndez y Jane Greenberg**

**Eva Méndez** es profesora titular del *Departamento de Biblioteconomía y Documentación* de la *Universidad Carlos III de Madrid*, donde es además directora del *Máster Universitario en Bibliotecas y Servicios de Información Digital*. Doctora en documentación, en los últimos 15 años ha realizado diversas actividades de docencia e investigación en temas relacionados con metadatos, web semántica, bibliotecas digitales, acceso abierto, políticas de información y web social. Es miembro del comité asesor de la *Iniciativa de metadatos Dublin core* (*DCMI*), co-coordinadora de la comunidad *DCMI Social Tagging,* y experta independiente para la *Comisión Europea* en bibliotecas digitales, y acceso abierto a datos de investigación.

*Universidad Carlos III de Madrid*
*Facultad de Humanidades, Comunicación y Documentación*
C/ Madrid, 128 (Dcho. 14.2.17) – 28903 Getafe (Madrid), España
*emendez@bib.uc3m.es*

**Jane Greenberg** es actualmente catedrática de excelencia en la *Universidad Carlos III de Madrid* en el *Departamento de Biblioteconomía y Documentación*. Es además profesora en la *School of Information and Library Science* (*SILS*) en la *University of North Carolina at Chapel Hill*, y directora del *Centro de Investigación de Metadatos* de la misma *Universidad*. Obtuvo su doctorado en la *University of Pittsburgh* y su máster, con la especialización de control bibliográfico, en la *Escuela de Biblioteconomía* de la *University of Columbia*. Ha recibido el premio *Margaret Mann Citation 2012* por su destacada labor profesional en el ámbito de la catalogación y la clasificación.

*Universidad Carlos III de Madrid – Facultad de Humanidades, Comunicación y Documentación*
C/ Madrid, 128 (Dcho. 14.2.53) – 28903 Getafe (Madrid), España
*School of Information and Library Science – University of North Carolina at Chapel Hill*
*janeg@email.unc.edu*

## Abstract

This paper summarizes new trends and advances in Knowledge Organization from the perspective of linked open data (LOD). Although this topic is particularly important to galleries, libraries, archives and museums, the so-called GLAM community, LOD is made significant by adoption beyond anyone community enabling a link to the global web. LOD includes descriptive metadata and vocabulary encoding schemes that are being "skosified" (encoded in the SKOS format) or rendered in OWL (the web ontology language) and made available not only "on" the web, but "for" the semantic web. The paper highlights a few exemplary initiatives in the field. The paper also introduces the *HIVE* (*Helping interdisciplinary vocabularies engineering*) framework and discusses the *HIVE-ES* (España) extension for Spanish language vocabularies, leading to a more global approach for linked open vocabularies (LOV).

## Keywords

Linked open vocabularies (LOV), Linked data (LD), Linked open data (LOD), Knowledge organization systems (KOS), Vocabularies, Metadata, SKOS, Semantic web, *HIVE*; *HIVE-ES*.

**Title: Datos enlazados para vocabularios abiertos y marco general de Hive**

## Resumen

Se presentan brevemente las nuevas tendencias y avances en la organización del conocimiento desde la perspectiva de *linked open data* (LOD). Aunque este tema es especialmente relevante para bibliotecas, archivos, museos y galerías –la llamada comunidad GLAM–, el interés de LOD se ha extendido a distintas comunidades que fomentan una mayor interacción con la web global. LOD incluye esquemas de metadatos descriptivos y de codificación de vocabularios que están siendo "skosificados" (codificados en el formato SKOS) o transformados en OWL (el lenguaje de ontologías web), accesibles no sólo "en" la Web, sino también "para" la web semántica. Se destacan algunas iniciativas paradigmáticas en este campo y se presenta el contexto general del proyecto *HIVE* (*Ayuda a la ingeniería de vocabularios interdisciplinares*) y se analiza la extensión *HIVE-ES* (España) para vocabularios en español, dando lugar a un enfoque más global de los vocabularios abiertos enlazados (LOV).

Nota: Este artículo puede leerse en su versión original en inglés en:
*http://www.elprofesionaldelainformacion.com/contenidos/2012/mayo/03_eng.pdf*

Artículo recibido el 19-05-12

**Palabras clave**

Vocabularios abiertos enlazados (LOV), Datos enlazados (LD), *Linked open data* (LOD), Sistemas de organización del conocimiento, Vocabularios, Metadatos, SKOS, Web semántica, *HIVE*, *HIVE-ES*

**Méndez, Eva**; **Greenberg, Jane**. "Linked data for open vocabularies and *HIVE*'s global framework". *El profesional de la información*, 2012, mayo-junio, v. 21, n. 3, pp. 236-244.

## 1. Introducción

En los últimos años se está prestando cada vez más atención a los datos enlazados (*linked data*), datos abiertos enlazados (*linked open data*) y a los vocabularios compartidos, en un contexto informativo cada vez más abierto y global. *Linked open data,* o simplemente LOD, se ha convertido en un término de moda que aparece en casi todas las iniciativas relacionadas con la organización de la información digital. LOD constituye una forma de gestionar la información que permitirá fortalecer y avanzar en la web semántica y la web de datos –una idea que cobró importancia al final de los 90 gracias al inventor de la World Wide Web, **Tim Berners-Lee**. En cierto sentido, *linked data* ha evolucionado como una norma *de facto* para publicar datos estructurados en la Web, involucrando a una gran variedad de colectivos que trabajan con distintos tipos de datos. La utilización de LOD está creciendo, de tal forma que cada vez son más las disciplinas comprometidas con buenas prácticas de *linked data* (LD) y con las tecnologías para dar a conocer y enlazar conjuntos de datos, posibilitando un acceso ilimitado a los mismos, compartiéndolos, integrándolos y reutilizándolos. En los últimos tiempos esta tendencia ha tenido un desarrollo asombroso, habiendo ya casi 3.500 conjuntos de datos publicados como LOD, según el *datahub* (*thedatahub.org*, mayo 2012).

> "Los vocabularios, formalizados como sistemas de organización del conocimiento (KOS), ayudan a abordar los problemas de la sobrecarga de información digital y favorecen la recuperación de información en internet"

Los vocabularios son la base de los datos enlazados. Formalizados como sistemas de organización del conocimiento (KOS), y basados en los lenguajes de un dominio, de una disciplina o comunidad, los vocabularios ayudan a abordar los problemas que subyacen a la sobrecarga de información digital y favorecen la recuperación de información en internet. Está claro que los vocabularios son importantes, no sólo en el mundo bibliotecario tradicional, sino también para muchos de los diferentes agentes implicados en la gestión de la información digital. En una encuesta reciente realizada por la compañía *Semantic Web*, en la que se entrevistaron a 158 participantes de distintas áreas como las tecnologías de la información, pertenecientes a diversos contextos científicos, del sector público o del ámbito educativo, se concluyó que el 85,5% utiliza vocabularios controlados en sus organizaciones. Una amplia proporción (88%) de estos usuarios de sistemas de organización del conocimiento opinaba, además, que muchas organizaciones pueden beneficiarse del uso de *linked data.* El 48,7% de los encuestados afirmó que los estándares como SKOS (*simple knowledge organization system*) son "muy importantes" y un 29,1% calificaron esta tendencia como "relevante" (**Kondert**; **Schandl**; **Blumauer**, 2011).

En el nuevo panorama de LD, ha aumentado el número de comunidades de usuarios que utilizan vocabularios, que tienen a su vez, enfoques diferentes y complementarios para compartirlos y aprovecharlos. Algunas de estas comunidades son:

- Desarrolladores web que incorporan dentro de sus páginas html datos estructurados que describen el contenido web. Para estas descripciones embebidas en el código fuente de un documento html utilizan estándares de codificación como microformatos, microdatos y RDFa (en vez de crear servicios Sparql), y usan un vocabulario de marcado compartido como *Schema.org*.
- La comunidad de la web semántica del *W3C*, que implementa y usa ontologías y vocabularios formalizados. Para ello utiliza OWL para construir dichos vocabularios y ontologías, y SKOS para crear KOS, como parte de un complejo proceso para el enriquecimiento de datos.
- Comunidades que crean y utilizan estándares de metadatos como el *Dublin core*, o estándares y perfiles de aplicación de metadatos particulares. Estos colectivos que crean sus estándares descriptivos recogen conjuntos de elementos y propiedades para conformar sus vocabularios, o esquemas (*schemas*) de metadatos en este caso.
- Usuarios de sistemas de organización del conocimiento (KOS) que abarcan estructuras de espacio-valor, como tesauros, clasificaciones de materias y archivos de autoridad, tradicionalmente considerados "vocabularios controlados".

Una visión interesante y sagaz del panorama LD la encontramos en un reciente post del blog de **Bernard Vatant** (2012), que recoge una interesante lista de instituciones que actualmente utilizan y publican vocabularios, o podrían potencialmente crearlos. Algunas de estas instituciones son por ejemplo: elaboradoras de estándares (*W3C*, *DCMI*); custodios/curadores del patrimonio institucional (*Library of Congress*) así como organizaciones globales que aglutinan el trabajo de los anteriores (ej. *IFLA*, *Europeana*, *OCLC*); grupos y asociaciones de medios (*BBC*, *The NYT*, *The Guardian*); proveedores de datos institucionales y gubernamentales (*data.gov*, *UN*, *World Bank*), centros de investigación (*DERI*, *Inria*); proyectos específicos de investigación; pequeñas y

medianas empresas del sector (*Talis*, *Mondeca*); grandes compañías de internet, a través de iniciativas como *Schema.org* (*Google*, *Bing*, *Yahoo!* and *Yandex*); así como otras iniciativas individuales o particulares de un dominio de información.

En cualquier caso, en LD se utilizan vocabularios de dos formas (al menos):

– como esquemas que proporcionan un conjunto de las propiedades que puede tener un objeto de información; y
– como esquemas de codificación de vocabularios que describen de manera formal el rango de valores que podría tener una propiedad concreta.

Los vocabularios enlazados ayudan a la adquisición de conocimiento a través de un control estricto y de una contextualización de los datos (conceptos, objetos, etc.). Este enfoque hace posible los procesos habituales de metadatos. A la vez, permite enlazar vocabularios con los registros de datos en sí mismos, proporcionando una infraestructura que mejora la eficacia del uso y recuperación de la información.

En primer lugar este artículo presenta una breve historia de las ideas que rodean al concepto de *linked data,* desde su fundamentación en la web semántica, hasta la evolución de las definiciones de datos enlazados (LD), datos abiertos enlazados (LOD), y vocabularios abiertos y enlazados (LOV). Seguidamente se hace una introducción al proyecto *HIVE* (*Helping interdisciplinary vocabularies engineering*), y a *HIVE-ES*, una extensión para los vocabularios en español, que contribuye a que la LOV tenga una dimensión más global. Finalmente se sintetizan las principales conclusiones y se ofrecen varias reflexiones que han surgido de este trabajo.

## 2. *Linked open* data en contexto

*Linked open data* (LOD) o *linked data* (LD) se ha convertido ya en un tema estándar de las convocatorias para la presentación de trabajos en el entorno de la biblioteconomía y documentación y de la informática. Asimismo se está prestando cada vez más atención a este tema (LD) en diferentes dominios disciplinares, como salud, gobierno y educación. De hecho, en casi todas las disciplinas que producen datos se puede identificar un grupo de defensores que promueven que los datos sean abiertos y que se puedan compartir. En este contexto, quizás de forma menos evidente, pero no por ello menos importante, se enmarca el interés de avanzar hacia vocabularios abiertos y enlazados (LOV), un paso más para la reutilización y compartición de los datos, y también para enlazar el conocimiento. Esta tendencia se hace eco de los argumentos propuestos anteriormente para desarrollar la web semántica; LD es clave para la evolución de la web semántica. De esta manera, en la prospección de LOV que hacemos aquí, es importante explicar y entender cómo encaja este nuevo desarrollo en la historia de la web.

### 2.1. Una breve historia de *linked data*

La idea de LD no surgió de la noche a la mañana. Al contrario, como ocurre con otros desarrollos en el ámbito de la documentación y de la informática, las ideas y los objetivos básicos son previos a la Web. Los ejemplos más comunes nos retrotraen a la noción de control bibliográfico universal, o a la presentación del *Memex* de **Vannevar Bush** (1945), un hipotético dispositivo que utilizaba enlaces asociativos para mantener la memoria a través del tiempo. Como afirmó recientemente **Dan Brickley** (2012), algunos de los trabajos actuales en LD y web semántica se pueden entender mejor si tenemos en cuenta el contexto de una larga historia, re-



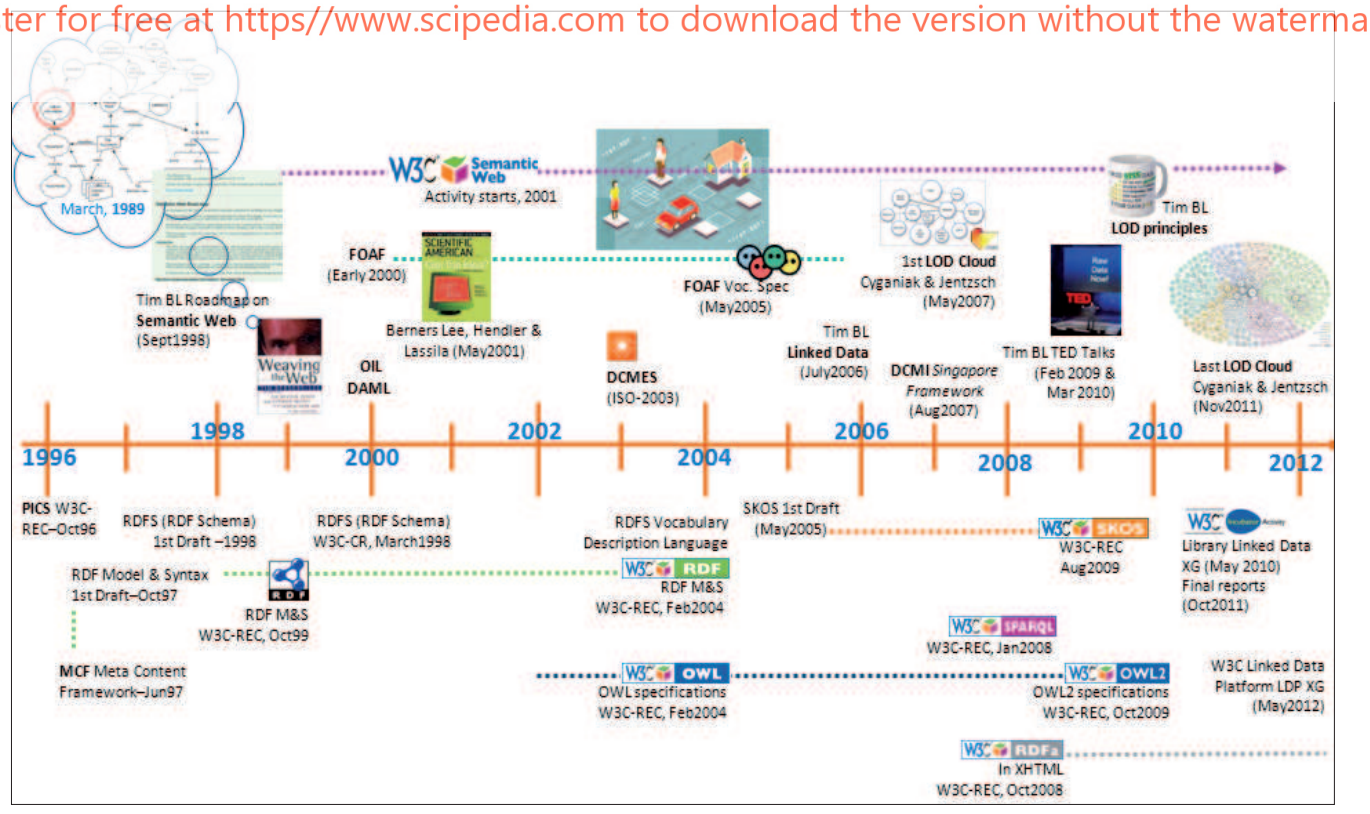


Figura 1. Evolución de la web semántica hacia *linked data* (Eva Méndez)

montándonos al informe anual de 1912 del *Instituto Belga de Bibliografía*.

Las ideas originales de **Tim Berners-Lee** no implicaban necesariamente la Web que conocemos hoy, pero apuntaban hacia el enlazado de conceptos y hechos –es decir, datos– a una escala global. Esta idea inicial de la web se ha difundido con el calificativo de web semántica, "una extensión de la Web actual en la que la información tienen un significado bien definido, permitiendo que los ordenadores y las personas trabajen mejor en cooperación" (**Berners-Lee**, 1999). **Berners-Lee** ha estado promocionando y defendiendo la misma idea por lo menos desde 1989: una potente estructura de conocimiento interconectado que enlaza información, documentos y datos. A esta idea le llamó primero la Web (1989), luego la web semántica (1989) y ahora *linked data* (2006). Tal y como predijo en su libro *Weaving the Web* (*Tejiendo la Web*) (1999), html hizo posible la web de documentos hipertextuales, y RDF y las tecnologías de la web semántica (OWL, SKOS, Sparql) harán posible la web de datos, a través de conjuntos de datos enlazados definidos en tripletes RDF.

La figura 1 muestra los hitos históricos más significativos en la construcción de las ideas de web semántica y LD, utilizando como hilo conductor los estándares del *WWW Consortium* (*W3C*) y las iniciativas que lo han hecho posible. Asimismo, se puede ver una recopilación exhaustiva de estos estándares de la web semántica, en español, en **Pastor** (2011).

La web semántica permite a las máquinas deducir significado a partir de datos estructurados que, publicados como *linked data*, pueden procesarse directamente. *Linked data* es un paso clave en la implementación de la web semántica, donde al representar las entidades de información a través de URIs, se convierten en procesables por máquina. Pero para alcanzar una web semántica real a nivel global, no es suficiente sólo enlazar datos o abrir datos. Es necesario que los datos sean tanto enlazados, como abiertos. Por eso **Berners-Lee**, dos décadas después de que inventó la web, y una década después de que fijara su atención en el desarrollo de la web semántica, exhortaba en la conferencia *TED* en 2009 a "¡abrir los datos en bruto, ahora!", y a expresarlos como datos enlazados, para poder sacarlos de sus silos.

## 2.2. Análisis y definición de *linked data* (LD)

Las explicaciones, argumentos y definiciones formales de *linked data* reflejan normalmente la fundamentación de **Berners-Lee** (2006) y reconocen LD como un conjunto de buenas prácticas para publicar y conectar datos estructurados en la Web. Implica la utilización de la Web para conectar datos relacionados que no estaban enlazados previamente, o para disminuir las barreras de uso de los datos que ahora están publicados utilizando otros métodos (*Linkeddata.org*, 2012). Toda esta fundamentación tiene básicamente cuatro reglas y cinco principios. Las reglas son bastante sencillas: 1) Utilizar URIs como nombres para las cosas (las URIs buenas para la web semántica son las que no cambian); 2) Utilizar URIs http (URIs *desreferenciables*) para que la gente pueda ver esos nombres; 3) Cuando alguien busca una URI, proporcionar información útil usando los estándares apropiados (RDF, Sparql); y 4) Incluir enlaces a otras URIs, de tal forma que se pueda recuperar más información. En 2010 **Berners Lee** redefinió su concepción de LD añadiendo la *filosofía de la apertura* de tal forma que los datos "abiertos" enlazados (LOD) se liberan bajo una licencia abierta, no impidiendo su libre reutilización, y usando la metáfora de las estrellas en los hoteles estableció cinco niveles o principios, para medir la calidad de LOD:

De manera general, LD puede entenderse como un enfoque para codificar datos con un gran nivel de detalle (granularidad). Los datos, en este contexto, pueden ser cualquier cosa (incluido un concepto, fijado en un término de un vocabulario), una declaración RDF o un conjunto de declaraciones que tengan un identificador en forma de URI. Los datos y los objetos que forman parte del paradigma de LD pueden provenir de un sistema de vocabulario normalizado como el *Dublin core* y de esquemas (*schemes*) de codificación en forma de vocabularios controlados, ontologías, taxonomías, ficheros de autoridad de nombres, sistemas de clasificación, etc. Los objetos se seleccionan a partir de estos vocabularios para la codificación de propiedades/valores o matizaciones de estos tipos de información. No hay límite para los tipos de vocabularios que pueden transformarse en LD. Se podría publicar como LD el *Diccionario Oxford* entero o el *Diccionario de la Real Academia Española*.

Los datos enlazados utilizan URIs (identificadores uniformes de recursos) únicos para cada tipo de recurso, de forma parecida a como se utilizan los identificadores en el ámbito bibliotecario para el control de autoridades. **Eric Miller** ha destacado la importancia de los identificadores y señala que su persistencia es una parte crucial para la integridad entre sistemas. La utilización de URIs http como forma de unificar las claves primarias locales dentro de una base de datos, permite trazar un espacio de datos universal, no sólo para las organizaciones bibliotecarias, sino para cualquier institución que quiera compartir información con otras organizaciones. "Tradicionalmente hemos mantenido estos identificadores locales dentro del sistema. Ahora estamos exponiendo los identificadores locales de tal forma que cualquier información externa puede engancharse a ellos" (**Miller**, 2011), así, todos los datos/cosas en el mundo de LD pueden enlazarse.

Son muchos los profesionales de la información y los investigadores que se han dado cuenta del valor añadido que



| | | |
|---|---|---|
| ★ | 1 | Disponible en la web (en cualquier formato) pero con una licencia abierta, para que sean datos abiertos |
| ★★ | 2 | Disponible como datos estructurados legibles por máquina (ej. excel en vez de una imagen escaneada de una tabla) |
| ★★★ | 3 | Igual que 2, pero añadiendo el principio de formato no propietario (ej. CSV en vez de excel) |
| ★★★★ | 4 | 1, 2 y 3, pero además utilizando estándares W3C (RDF, Sparql) para identificar las cosas de tal forma que se puedan apuntar |
| ★★★★★ | 5 | 1, 2, 3 y 4 pero además, enlazar tus datos con los de otros para proporcionar contexto |

Figura 2. Principios de la apertura de datos añadidos al paradigma *linked data* (**Berners-Lee**, 2006 actualizado en 2010)

pueden obtener bibliotecas, archivos, museos y galerías, abriendo y enlazando sus datos culturales. De manera especial podemos destacar los trabajos de **Byrne** y **Goddard** (2011) en Canadá, **Oomen** *et al.* en Holanda (2012), **Saorín** (2012) y **Peset** *et al.* (2011) en España, y **Jon Voss** (LOD-LAM, 2011) en Estados Unidos. Pero la contribución más importante, desde el punto de vista de las organizaciones de la memoria y en el área de las humanidades digitales son los informes del *Grupo incubadora del W3C sobre datos bibliotecarios enlazados* (*W3C LLD XG*, 2011a y b; **Baker**, 2012). Este grupo define los datos enlazados como datos publicados de acuerdo con los principios que facilitan la vinculación entre ellos, conjuntos de elementos y vocabularios de valor. El enfoque de *HIVE,* que trataremos después, se centra en estos últimos esquemas de codificación de vocabularios.

## 3. De LOD a LOV: vocabularios abiertos enlazados como parte del nuevo ecosistema de la organización del conocimiento

Distintas investigaciones han demostrado que la búsqueda por materias o la búsqueda conceptual basada en un tema es la forma más común de buscar en la Web (**Yu**; **Young**, 2004; **Savolainen**; **Kari**, 2006). Es decir, la gente busca más frecuentemente información sobre un tema, por ejemplo, un lugar para viajar, una estado de salud, un evento histórico, etc., que una organización específica, el resumen oficial de una nueva película o un artículo en particular. Incluso cuando busca una "entidad conocida" como una persona o un lugar se pueden usar términos de carácter temático o conceptual. Buscamos contenidos por materias, y por eso necesitamos una búsqueda semántica.

> "Buscamos contenidos por materias, y por eso necesitamos una búsqueda semántica"

Por ejemplo una persona que quiere encontrar información sobre la historia del Parque del Retiro en Madrid puede iniciar su consulta web por los conceptos "España" y "parques", y quizás por el nombre "Retiro". Probablemente, la persona que hace esta búsqueda está más interesada en la ampliación del parque en 1632 como un refugio fuera de la muralla de la ciudad por el rey Felipe IV y la familia Real, o cómo se utilizaba el parque durante la guerra civil española, que en averiguar sus coordenadas geográficas.

El contenido temático puede representarse a través de conceptos de vocabularios *skosificados*. Con vocabularios abiertos y enlazados se despliega un enorme potencial para aprovecharlos en el ámbito global de la Web, vinculando la información en un entorno abierto.

### 3.1. Abrir y enlazar vocabularios para crear un nuevo ecosistema de sistemas de organización del conocimiento

En este nuevo ecosistema, los vocabularios deben publicarse como LOD para poderse reutilizar. Hay muchos vocabularios disponibles en la Web aunque con una codificación diferente, por ejemplo, en html. Hace algún tiempo señalamos el potencial de XML/RDF para codificar y representar vocabularios en un entorno de semántica controlada como un nivel corporativo o intranet (**Méndez**, 2000). Sin embargo, el desarrollo de la web semántica basada en estándares (fig. 1) abre la posibilidad de una interoperabilidad más global a través de la web de datos enlazados. SKOS, como estándar recomendado del *Consorcio Web* desde 2009, implica un paso más para hacer los tesauros interoperables, permitiendo que se puedan compartir y poniéndolos a disposición de tal forma que su contenido se puede integrar y vincular. Este enfoque de la skosificación de vocabularios impulsa la propia idea de web semántica, donde los vocabularios se crean explícitamente para la web, y es donde realmente tiene sentido el concepto de vocabularios abiertos y enlazados que recogemos aquí.

La implementación técnica para enlazar vocabularios abiertos se basa en tecnologías, estándares y buenas prácticas de la web semántica, tales como RDF/SKOS, OWL servicios Sparql y almacenamiento de conceptos definidos en tripletas. Estas tecnologías de la web semántica ofrecen un gran potencial para una mejor diseminación e integración de datos en otros entornos y aplicaciones (**Vatant**, 2012). Con los estándares de la web semántica no sólo ponemos vocabularios "en" la web, sino "para" la web semántica, permitiendo a las máquinas utilizarlos directamente. El *Datahub* (lo que conocemos como la nube de *linked open data*) es un catálogo elaborado de forma colectiva que recoge distintos conjuntos de datos, que incluye, entre la enorme cantidad de "cosas" publicadas actualmente como LD, también esquemas (*schemes*) de codificación de vocabularios. LOV puede entenderse como un subconjunto clarificado de la desordenada y compleja nube de LOD, donde incluimos aquellos vocabularios que pueden considerarse sistemas de organización del conocimiento, comparables a otros estándares. **Vatant** (2012) en su detallado y sugerente post titulado *LOV stories* dice que "hay tantos datos buscando buenos vocabularios como buenos vocabularios buscando datos". Piensa además que las actividades para crear un nuevo ecosistema para los vocabularios abiertos y enlazados deben basarse en la filosofía del procomún y de la *co-opetition*: crear vocabularios como recursos compartidos en los que todos los agentes implicados tienen el mismo interés, de tal forma que se establece una competición cooperativa.

> "La utilización de estándares de la web semántica permite no sólo poner vocabularios *en* la web, para que la gente los pueda leer, sino también publicarlos *para* la web semántica, para que las máquinas los usen directamente"

### 3.2. Trazando el nuevo panorama de los sistemas de organización del conocimiento. Iniciativas de vocabularios abiertos y enlazados

Existen varios registros colectivos que proporcionan acceso a vocabularios abiertos en el nuevo entorno de *linked data*

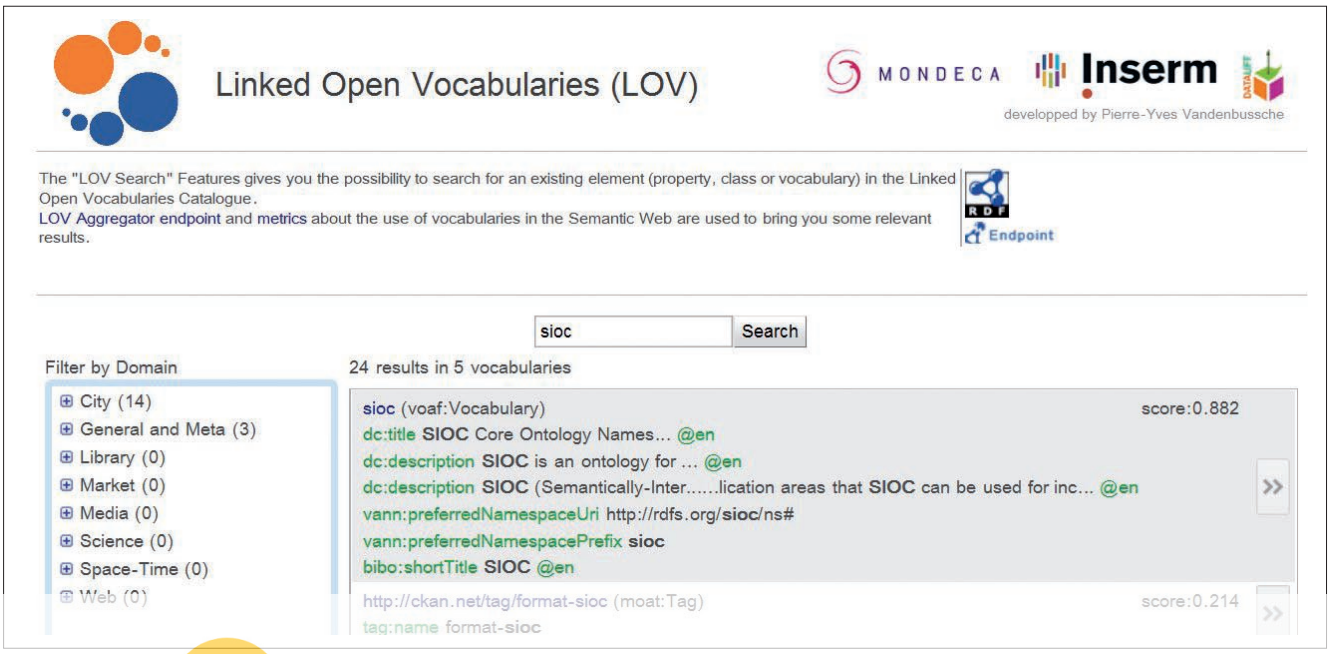


Figura 3. *Proyecto LOV*, búsqueda del vocabulario *SIOC* (*Semantically-interlinked online communities*)

y web semántica, reconfigurando un nuevo paradigma para los sistemas de organización del conocimiento en red. Destacamos a continuación cuatro de estas iniciativas:

***Linked open vocabularies* (*LOV*).** El registro *LOV*, o proyecto *LOV*, creado por **Bernard Vatant** y **Pierre-Yves Vandenbussche** y publicado por *Mondeca Labs* como parte del proyecto *Datalift* en marzo de 2012, se ha presentado a la *Open Knowledge Foundation* (*OKFN*), para solicitar ser alojado en los servidores de la *OKFN* como ocurre con el *Datahub*. Tal y como se establece en tal candidatura, el proyecto *LOV* tiene la finalidad de proporcionar un acceso fácil a vocabularios, en particular a la forma en que se enlazan unos con otros y también pretende analizar a través de métricas cómo se utilizan en la nube de *linked data* para ayudar a mejorar su comprensión, visibilidad, usabilidad y calidad general.

*LOV* es uno de los proyectos más innovadores en esta área, y como ya hemos señalado, muchas de sus ideas legitiman las que reflejamos en este artículo. *LOV* presenta un "ecosistema cada vez mayor de vocabularios abiertos enlazados (RDFs u ontologías OWL) enlazados en la nube LOD".

El catálogo *LOV* da acceso a 262 vocabularios, representados en RDF, OWL y SKOS, que se clasifican por áreas, y se interrelacionan utilizando un vocabulario específico denominado VOAF (*vocabulary of a friend*) para describir ontologías que forman parte de la nube de *linked data.* Funciona básicamente como un registro, pero da un paso más al clasificar los vocabularios desde una perspectiva más amplia.

***Open metadata registry* (*OMR*)**. Esta iniciativa "comenzó como el registro de la *NSDL* (*National Science Digital Library*) de los EUA tratando de responder a la pregunta: ¿qué deberían hacer estos registros y cómo deberían funcionar en un entorno de servicios abiertos?" (**Phipps**; **Hillman**, 2011) y, con el tiempo, se ha convertido en el registro más importante para la comunidad bibliotecaria. *OMR* transciende a la comunidad de bibliotecas, archivos y museos, ya que tiene su origen en la educación científica, e incluye una gran variedad de vocabularios (conjuntos de elementos, ontologías y también vocabularios controlados). Un aspecto innovador de este registro es su *sandbox,* un espacio virtual de pruebas donde cualquier usuario puede experimentar, jugar, compartir y aprender cómo participar en el entorno de LOV. Se está preparando que *OMR* soporte vocabularios multilingües y versiones lingüísticas distintas pero asociadas. Una de sus principales creadoras, **Diane Hillman**, es la fundadora y líder de la *Comunidad Dublin core sobre Gestión de Vocabularios* (*DCMI-VMC*), un foro que se encarga de analizar buenas prácticas en el ámbito de los vocabularios en la web semántica.

***Amalgame* (*Amsterdam alignment generation metatool*)** es un servidor interactivo de alineación de vocabularios, aún en desarrollo, realizado por la *Universidad de Amsterdam*, que surgió en el contexto de los proyectos europeos *Presto Prime* y *Europeana Connect*. Implica un paso más en la dinámica de trabajo para skosificar vocabularios y convertir colecciones de metadatos al *Modelo de datos de Europeana* (*EDM*). *Amalgame* tiene como objetivo encontrar, evaluar y gestionar la correspondencia de términos entre vocabularios dentro del contexto de la *Ontology Alignment Evaluation Initiative* (*OAEI*). En *OAEI* se establecen diferentes métodos de búsqueda de correspondencias, que pueden combinarse utilizando una organización específica de los flujos de trabajo. La principal diferencia que presenta *Amalgame* en relación con el proyecto *LOV* es que *Amalgame* incluye un número limitado de vocabularios, relacionados fundamentalmente con el patrimonio cultural, con correspondencias entre ellos. Y la principal diferencia con respecto a OMR es que *Amalgame* sólo incluye esquemas de codificación de vocabularios, no esquemas o conjuntos de elementos y propiedades, y busca una alineación entre los distintos términos incluidos en dichos vocabularios, basándose en vocabularios eje.

***NCBO-BioPortal*** es una aplicación web que da acceso a 302 ontologías biomédicas y vocabularios, incluyendo tesauros,

en el campo de la biomedicina y la biología. Se puede navegar, buscar y descargar ontologías. También existe la posibilidad de mapeo entre vocabularios, para lo cual las ontologías registradas se codifican fundamentalmente en OWL, u OBO, un lenguaje específico para la representación de ontologías en el dominio de las bio-ciencias. También, cada vez más, se registran y proporcionan representaciones en SKOS. Todos estos vocabularios y ontologías se utilizan para la búsqueda conceptual de recursos biomédicos. *NCBO-BioPortal* no es sólo un servicio de vocabularios especializados, una vez que un término buscado se refina y se alinea entre diferentes vocabularios, se puede hacer clic y explorar los recursos sobre dicho tema en su sitio original. La principal diferencia de este servicio con respecto a los registros de vocabularios descritos anteriormente es su enfoque vertical de un dominio específico, en vez de registrar de manera integral vocabularios formalizados de cualquier disciplina.

> *HIVE* y *HIVE-ES* proporcionan una base sólida para enlazar y abrir vocabularios interdisciplinares en un entorno de información multilingüe

## 4. Marco *HIVE-ES*: skosificando, abriendo y enlazando vocabularios

*HIVE* (*Helping interdisciplinary vocabulary engineering*) es un proyecto iniciado con el apoyo del *Institute of Museum and Library Services* (*IMLS*) de Estados Unidos. *HIVE* presenta un modelo para la utilización de vocabularios enlazados abiertos para crear metadatos de materia de forma dinámica en el momento de la indización, extrayéndolos a partir de múltiples vocabularios. Este proceso posibilita una selección de los conceptos más adecuados para representar el contenido de un recurso, que no se limita a un único vocabulario. En la metáfora que utilizan **Greenberg** y los miembros del equipo de *HIVE* para explicar este proyecto, es como si una abeja fuera a distintas flores (vocabularios) buscando polen y llevara pequeñas pepitas de polen de vuelta al panal (sistema de información digital); en este caso, esas pepitas de polen son los distintos conceptos relevantes para la indización. Nos hemos dado cuenta de que esta metáfora es compatible con la de **Vatant** (2012), donde en sus "jardines LOV", los distintos vocabularios como *Schema.org* son árboles o flores de un jardín.

La iniciativa *HIVE-ES* (España) se fundamenta en el trabajo previo del proyecto *HIVE* para abordar los retos de los sistemas de organización del conocimiento y proporcionar una forma de buscar y generar simultáneamente metadatos de materia, extrayéndolos de múltiples vocabularios en idioma español. Concretamente, *HIVE* y *HIVE-ES* son dos proyectos que se han iniciado con el objetivo de abordar los problemas que conllevan los vocabularios controlados en relación al coste, la interoperabilidad, y las limitaciones de usabilidad (**Greenberg** *et al.*, 2011):

- Los vocabularios controlados son caros de crear y mantener.
- Su adhesión o conformidad con estándares *ISO*, *ANSI/NISO*, *W3C*, *IETF* u otros estándares, no corrobora la interoperabilidad de sistemas de organización del conocimiento (KOS).
- El diseño de sistemas de organización del conocimiento no siempre implica un fácil acceso y uso de los mismos.

La iniciativa *HIVE-ES* amplía la actividad inicial de *HIVE* para abordar estos retos en los países de habla hispana, que producen información y sistemas de organización del conocimiento en español. *HIVE-ES* permite la búsqueda y generación simultánea de metadatos orientados al contenido, extrayendo los términos de indización de múltiples vocabularios en español. Este proyecto se ha lanzado en España, en el seno del grupo de investigación *Tecnologías Aplicadas a la Información y la Documentación*) del *Departamento de Biblioteconomía y Documentación* de la *Universidad Carlos III de Madrid*, la *Biblioteca Nacional de*




Figura 4. Wiki del proyecto *HIVE-ES*

*España* (*BNE*), y el *Centro de Investigación de Metadatos* de la *School of Information and Library Science* (*SILS*) en la *University of North Carolina at Chapel Hill* (*SILS-MRC*).

*HIVE-ES* está construyendo vocabularios y proporciona una demostración para la indización interdisciplinar a partir de diversos vocabularios en español. Los vocabularios integrados en *HIVE-ES* están fundamentalmente creados como *linked data,* utilizando SKOS como lenguaje de codificación para representar sus clases y propiedades.

El servidor de vocabularios *HIVE-ES* tiene tres vocabularios (mayo 2012): la sección española de *Agrovoc* –el vocabulario de la *FAO* (*United Nations Food and Agriculture*)-, la *Lista de encabezamientos de materia* (*LEM*), y los *Encabezamientos de materia de la Biblioteca Nacional de España (EmBNE),* que es el equivalente español de la *Lista de Encabezamientos de Materia* de la *Library of Congress* de los EUA. La integración de los *EmBNE* fue la tarea más compleja, teniendo en cuenta la necesidad de convertir los datos en formato Marc21 de autoridades a SKOS. El proceso implicó varias etapas: 1) Mapeo conceptual de los campos Marc21 (ej., 1XX, 4XX y 5XX) a las correspondientes etiquetas de SKOS (ej., 'skos:prefLabel' y 'skos:altLabel'); 2) lectura y análisis gramatical (*parsing*) del archivo *EmBNE* en formato Marc21; y 3) reetiquetado de los contenidos individuales de la *EmBNE* a las correspondientes en SKOS. La conversión no se apoya en una sola herramienta, aunque sí fue de gran ayuda la librería java *Marc4J,* que define una API para analizar grandes archivos Marc.

En cualquier caso, el marco de trabajo y la infraestructura de *HIVE-ES* están en consonancia con el proyecto *HIVE* original, que incluye:

- *HIVE core* para la extracción automática de metadatos, detección de temas y recuperación de contenidos. Para la extracción se utiliza un algoritmo de extracción por frases clave denominado KEA (*keyphrase extraction algorithm*); y la recuperación de conceptos está basada en *Lucene.* Aunque estas son las principales funciones del sistema, *HIVE core* incluye además el almacenamiento y gestión de tripletes RDF utilizando *Elmo* para almacenar objetos y propiedades en un repositorio RDF implantado en *Sesame*.
- El servicio de vocabularios *HIVE* incluye la interfaz web que permite a los usuarios navegar y buscar en los distintos vocabularios incluidos en el sistema demo. *HIVE* desarrolló este servicio de vocabularios utilizando *Google web toolkit,* que ahora se traslada a *HIVE-ES*.
- El componente *HIVE REST* (*representational state transfer*) proporciona una API basada en servicios web para facilitar la integración de otros componentes de software externos que pudieran surgir.

*HIVE-ES* está en sus comienzos de desarrollo e implementación, por ello, tenemos planificado seguir skosificando e incluyendo más vocabularios en español. Las implementaciones de *HIVE* documentadas hasta ahora son monolingües, y el código actual de momento no soporta la integración de vocabularios multilingües. Pero la naturaleza abierta de esta iniciativa y el código compartido abre una puerta a otros desarrolladores y posibilidades que puedan contribuir a *HIVE* y *HIVE-ES* en esa línea. La dimensión española del proyecto *HIVE* instiga estos desarrollos, poniendo de relieve la necesidad de contemplar vocabularios en otros idiomas distintos al inglés, que permitan avanzar hacia un contexto más global e integrador, en un universo de información web necesariamente multilingüe.

## 5. Conclusiones

Este artículo presenta una visión de las diferentes tendencias y enfoques que se están llevando a cabo para el avance de la organización del conocimiento, y de vocabularios compartidos, desde la perspectiva de *linked data* (LD) y *linked open data* (LOD). Resumimos a continuación, algunas de las conclusiones más sólidas que se desprenden de este trabajo.

- La infraestructura tecnológica que soporta LD/LOD es ya bastante potente gracias al uso de las tecnologías en red y el desarrollo y la adopción de los estándares del *W3C* (RDF, SKOS, Sparql, etc.).
- Los vocabularios controlados, en sus múltiples formas (tesauros, taxonomías, ontologías y lenguajes disciplinares o vinculados a un dominio o comunidad), pueden aumentar su potencial de implementación a través de su codificación en RDF/SKOS.
- Son muchas las comunidades que están adoptando este potencial, skosificando sus vocabularios, haciéndolos más abiertos e interoperables.
- Los registros de vocabularios están avanzando hacia un nuevo nivel, con desarrollos y proyectos como los que hemos destacado en este artículo (proyectos *LOV*, *OMR*, *Amalgame* y *NCBO BioPortal*). Estos registros proporcionan formas sostenibles de compartir vocabularios y permiten diversas operaciones en el ámbito de la web semántica.
- En la publicación de vocabularios abiertos no basta sólo con desarrollar vocabularios *para* la web. Los principios de las "cinco estrellas" de **Tim Berners-Lee** son también aplicables para que los sistemas de organización del conocimiento incrementen sus posibilidades de almacenamiento de información, para su uso efectivo (licencias abiertas).
- La comunidad de usuarios y organizaciones que adoptan LOD y LOV está creciendo rápidamente y con entusiasmo, abarcando tanto dominios científicos, como iniciativas impulsadas en el ámbito de las humanidades (p. ej., *Europeana*).

Los tesauros propietarios, y otros sistemas de organización del conocimiento, pueden formar parte de este nuevo panorama. De esta forma se están produciendo diversas iniciativas para avanzar más allá de su mera visualización web de carácter textual o hipertextual. Estos sistemas de organización del conocimiento (KOS) están aumentando tanto "en" la web como "para" la web semántica. Los ejemplos contemplados en este artículo, así como otros aspectos de la infraestructura de *linked data,* constituyen importantes avances hacia este nuevo ecosistema. *HIVE* constituye sólo un modelo en este nuevo panorama de LOV. Por su parte, *HIVE-ES*, la extensión del *HIVE* original para los vocabularios en idioma español, introduce un nuevo enfoque para

los vocabularios abiertos y enlazados. Y todo ello establece una sólida fundamentación para la apertura y el enlazado de vocabularios en un entorno multilingüe, donde necesitamos compartir vocabularios de espacio-valor entre diversos idiomas.

Teniendo en cuenta que los proyectos de información digital aspiran a que sus recursos, o las descripciones de sus colecciones de recursos, sean accesibles a través de la web global, se están haciendo varios esfuerzos de investigación de naturaleza interdisciplinar. Los vocabularios abiertos y enlazados son parte de esta tendencia creciente, y surgen iniciativas que permiten la búsqueda y la indización, a través de múltiples vocabularios, que necesitan que se siga trabajando para alcanzar un entorno completamente interoperable. Los avances para poder compartir estas nuevas formas de representación de la información necesitan una infraestructura social y tecnológica mayor, que permita aunar personas, tecnologías y vocabularios. Además, para que este enfoque sea sostenible es clave que exista una visión compartida y un sentido de comunidad. La *Comunidad de Gestión de Vocabularios* de la *DCMI* está cobrando un especial interés en este sentido, gracias al liderazgo, la inspiración y el compromiso de miembros de los proyectos *OMR* y *LOV*. La recopilación y coordinación de proyectos y desarrollos, y el trabajo comunitario permitirán que esta evolución, necesaria en el ámbito de la organización del conocimiento y los vocabularios, alcance su máximo potencial.

## 6. Agradecimientos

Agradecemos a los colegas **Dan Brickley** y **Charles McCathie-Neville** la cuidadosa revisión de este artículo, así como sus valiosos comentarios y retroalimentación. Asimismo queremos agradecer a los miembros del equipo de *HIVE* y *HIVE-ES* todas sus contribuciones al proyecto. Los fundamentos de la tecnología *HIVE* están apoyados por la *IMLS* a través de la financiación del proyecto *LG-07-08-0120-08*.

## 7. Referencias

**Baker, Thomas**. "Libraries, languages of description, and linked data: a Dublin core perspective". *Library hi tech*, 2012, v. 30, n. 1, pp. 116-133.
*http://dx.doi.org/10.1108/07378831211213256*

**Berners-Lee, Tim**. *Information management: a proposal*, 1989.
*http://www.w3.org/History/1989/proposal.html*

**Berners-Lee, Tim**. "Semantic web road map". *Design issues*, 1998, September.
*http://www.w3.org/DesignIssues/Semantic.html*

**Berners-Lee, Tim**. *Weaving the Web: The original design and ultimate destiny of the World Wide Web by its inventor.* San Francisco: Harper, 1999.

**Berners-Lee, Tim**. "Linked data". *Design issues*, 2006-07-27.
*http://www.w3.org/DesignIssues/LinkedData.html*

**Berners-Lee, Tim**. "The next Web of open, linked data". *TED*, February, 2009
*https://www.ted.com/talks/tim_berners_lee_on_the_next_web.html*

**Brickley, Dan**. "Semantic web, part 5". In: *Libraries, media & the semantic web*, hosted by the *BBC*, 28 March 2012, video.
*http://www.youtube.com/watch?v=_-6mhdjE1XE&feature=player_embedded*

**Bush, Vannevar**. "As we may think". *The Atlantic monthly*, 1945, July.
*http://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/3881*

**Byrne, Gillian**; **Goddard, Lisa**. "The strongest link: Libraries and linked data". *DLib magazine*, 2010, v. 16, n. 11-12.
*http://www.dlib.org/dlib/november10/byrne/11byrne.html*

**Greenberg, Jane** *et al.* "*HIVE*: Helping interdisciplinary vocabulary engineering". *Bulletin of the American Society for Information Science and Technology*, 2011, v. 37, n. 4.
*http://www.asis.org/Bulletin/Apr-11/AprMay11_Greenberg_etAl.html*

**Kondert, Florian**; **Schandl, Thomas**; **Blumauer, Andreas**. *Do controlled vocabularies matter? Survey results.* Semantic Web Company, Viena, 2011.
*http://issuu.com/andreas_blumauer/docs/survey_do_controlled_vocabularies_matter_2011_june#download*

**Méndez, Eva**. "Metadatos y tesauros: aplicación de XML/RDF a los sistemas de organización del conocimiento en intranets". In: *VII Jornadas españolas de documentación*, Bilbao (Spain), 2000, October, 19-20, pp. 211-219.
*http://hdl.handle.net/10760/12698*

**Miller, Eric**. "Linked data and libraries" [recorded by Micheline Westfall]. *Serials librarian*, 2011, v. 60, n. 1-4, pp. 17-22.
*http://dx.doi.org/10.1080/0361526X.2011.556427*

**Oomen, Johan**; **Baltussen, Lotte-Belice**; **Van Erp, Marieke**. "Sharing cultural heritage the linked open data way: why you should sign up". In: *Museums and the Web 2012*, San Diego, April, 11-14.
*http://www.museumsandtheweb.com/mw2012/papers/sharing_cultural_heritage_the_linked_open_data*

**Pastor, Juan-Antonio**. *Tecnologías de la web semántica.* Colección EPI. Barcelona: UOC, 2011.

**Peset, Fernanda**; **Ferrer-Sapena, Antonia**; **Subirats-Coll, Imma**. "Open data y linked open data: su impacto en el área de bibliotecas y documentación". *El profesional de la información*, 2011, marzo-abril, v. 20, n. 2, pp. 165-173.
*http://dx.doi.org/10.3145/epi.2011.mar.06*

**Phipps, Jon**; **Hillman, Diane**. "The Open Metadata Registry: an update". *Bulletin of the American Society for Information Science and Technology*, 2011, v. 37, n. 4.
*http://www.asis.org/Bulletin/Apr-11/AprMay11_Phipps_Hillmann.pdf*

**Saorín, Tomás**. "Cómo linked open data impactará en las bibliotecas a través de la innovación abierta". *Anuario ThinkEPI*, 2012, v. 6, pp. 288-292.
*http://hdl.handle.net/10760/16913*

**Savolainen, Reijo**; **Kari, Jarkko**. "User-defined relevance criteria in web searching". *Journal of documentation*, 2006, v. 62, n. 6, pp. 685-707.
*http://dx.doi.org/10.1108/00220410610714921*

**Vatant, Bernard**. LOV stories, Part 2: Gardeners and gatekeepers. *The wheel and the hub: Tracks in the knowledge commons* [blog], 22-03-2012.
*http://blog.hubjects.com/2012/03/lov-stories-part-2-gardeners-and.html*

W3C LLD XG. *Library Linked Data Incubator Group Final Report*, 25 October 2011. Tomas Baker *et al.*, eds., 2011a.
*http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025*

W3C LLD XG. *Library Linked Data Incubator Group: Datasets, value vocabularies, and metadata element sets*, 25 October 2011, Antoine Isaac *et al.*, eds., 2011b.
*http://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset-20111025*

**Yu, Holly**; **Young, Margo**. "The impact of web search engines on subject searching in opac". *Information technology and libraries*, 2004, v. 23, n. 4, pp. 168-180.
*http://dx.doi.org/10.1234/12345678*

## 8. Webs relevantes citadas

Amalgame (Amsterdam alignment generation metatool)
*http://semanticweb.cs.vu.nl/amalgame*

DC, DCMI (Dublin core, Dublin core metadata initiative)
*http://dublincore.org*

DCMI-VMC (DCMI Vocabulary Management Community)
*http://dublincore.org/groups/vocabulary-management*

DataHub
*http://thedatahub.org*

EDM (Europeana data model)
*http://pro.europeana.eu/edm-documentation*

FOAF (friend of a friend)
*http://www.foaf-project.org*

HIVE (Helping interdisciplinary vocabularies engineering)
*http://ils.unc.edu/mrc/hive*

HIVE Project Wiki
*https://www.nescent.org/sites/hive/Main_Page*

HIVE-ES (Helping interdisciplinary vocabularies engineering-España). Project wiki
*http://klingon.uc3m.es/hive-es/wiki*

HIVE-ES server
*http://klingon.uc3m.es:8080/home.html*

KEA (keyphrase extraction algorithm)
*http://www.nzdl.org/Kea*

Linked Data
*http://linkeddata.org*

LOV project (linked open vocabularies)
*http://labs.mondeca.com/dataset/lov*

Marc4J java library
*http://marc4j.tigris.org*

NCBO-BioPortal
*http://bioportal.bioontology.org*

OAEI (Ontology alignment evaluation initiative)
*http://oaei.ontologymatching.org*

OBO (The open biological and biomedical ontologies)
*http://obofoundry.org*

OMR (Open metadata registry)
*http://metadataregistry.org*

OKFN (Open Knowledge Foundation)
*http://okfn.org*

Schema.org
*http://schema.org*
*http://blog.schema.org*

Semanticweb.com
*http://semanticweb.com*

SIOC (Semantically-interlinked online communities)
*http://sioc-project.org*

VOAF (Vocabulary of a friend)
*http://labs.mondeca.com/vocab/voaf*

W3C (World Wide Web Consortium)
- Linked data
  *http://www.w3.org/standards/semanticweb/data*
- MCF (meta content framework)
  *http://www.w3.org/TR/NOTE-MCF-XML-970606*
- Sparql
  *http://www.w3.org/TR/rdf-Sparql-query*
- RDF (resource description framework)
  *http://www.w3.org/standards/techs/rdf*
- RDFa (RDF attributes)
  *http://www.w3.org/TR/xhtml-rdfa-primer*
  *http://rdfa.info*
- RDFS (RDF schema)
  *http://www.w3.org/TR/rdf-schema*
- SKOS (simple knowledge organization system)
  *http://www.w3.org/standards/techs/skos*
- Semantic web
  *http://www.w3.org/standards/semanticweb*
- Vocabularies
  *http://www.w3.org/standards/semanticweb/ontology*